The Journal of Politics

Big Data Justice: A Case for Regulating the Global Information Commons

--Manuscript Draft--

Manuscript Number:	182612R2
Full Title:	Big Data Justice: A Case for Regulating the Global Information Commons
Article Type:	Research Article
Section/Category:	Political Theory
Corresponding Author:	Kai Spiekermann, Ph.D. London School of Economics and Political Science London, London UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	London School of Economics and Political Science
Corresponding Author's Secondary Institution:	
First Author:	Kai Spiekermann, Ph.D.
First Author Secondary Information:	
Order of Authors:	Kai Spiekermann, Ph.D.
	Adam Slavny, PhD
	David V Axelsen, PhD
	Holly Lawford-Smith, PhD
Order of Authors Secondary Information:	
Abstract:	The advent of artificial intelligence (AI) challenges political theorists to think about data ownership and policymakers to regulate the collection and use of public data. AI producers benefit from free public data for training their systems while retaining the profits. We argue against the view that the use of public data must be free. The proponents of unconstrained use point out that consuming data does not diminish its quality and that information is in ample supply. Therefore, they suggest, publicly available data should be free. We present two objections. First, allowing free data use promotes unwanted inequality. Second, contributors of information did not and could not anticipate that their contribution would be used to train AI systems. Therefore, charging for extensive data use is pro tanto permissible and desirable. We discuss policy implications and propose a progressive data use tax to counter the inequality arising.

Big Data Justice:
A Case for Regulating the Global Information Commons

Short title: Big Data Justice

Kai Spiekermann

London School of Economics, Department of Government, Houghton Street, London, WC2A 2AE, UK. Email: k.spiekermann@lse.ac.uk

Adam Slavny

University of Warwick, School of Law, Coventry, CV4 7AL, UK. Email: a.slavny@warwick.ac.uk

David V. Axelsen

University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK. Email: d.v.axelsen@essex.ac.uk

Holly Lawford-Smith[1]

University of Melbourne, School of Historical and Philosophical Studies, Faculty of Arts, Melbourne, Victoria 3010, Australia. Email: holly.lawford-smith@unimelb.edu.au

[1] This article was written by The London Cooperative, a philosophy writing collective made up of the four authors. All authors contributed equally.

Abstract

The advent of artificial intelligence (AI) challenges political theorists to think about data ownership and policymakers to regulate the collection and use of public data. AI producers benefit from free public data for training their systems while retaining the profits. We argue against the view that the use of public data must be free. The proponents of unconstrained use point out that consuming data does not diminish its quality and that information is in ample supply. Therefore, they suggest, publicly available data should be free. We present two objections. First, allowing free data use promotes unwanted inequality. Second, contributors of information did not and could not anticipate that their contribution would be used to train AI systems. Our argument implies that managing the 'global information commons' and charging for extensive data use is permissible and desirable. We discuss policy implications and propose a progressive data use tax to counter the inequality arising.

Keywords: big data, artificial intelligence, machine learning, data ownership, inequality

Financial Support: David Axelsen gratefully acknowledges support from Det Frie Forskningsråd (grant DFF – 4089-00313).

After decades of incremental growth, progress in the field of artificial intelligence (AI) is suddenly exploding. Better hardware and machine learning algorithms certainly play a role in this acceleration. But another crucial factor is the availability of 'big data' on a scale the world has not seen before (Polson and Scott 2018, 6). Remarkably, while the producers of AI draw on a public good – the reams of data provided by millions of contributors to the public parts of the World Wide Web – they typically create private goods: software or machines with proprietary code. For example, navigation apps collect huge datasets from public and non-public sources to improve their services[2] and automated translation services can be trained using publicly available bilingual documents (e.g., Koehn 2010). This raises interesting normative questions. First, why should the profits reaped from these systems be privately retained, given that much of the data was provided by the public? And second, how should policymakers regulate the use of publicly available data? Some maintain that information and data should always be free to use. They justify this view by pointing out that information does not diminish in quality, regardless of how many people use it. They also observe that the information available on the World Wide Web is in rich abundance, so that a tax to overcome free-riding is unnecessary. The central aim of this paper is to show that this 'Free Use Argument' is unsound. This, in turn, suggests that societies are not obligated to allow unrestricted and unrecompensed use of publicly available information and that different ways to limit or charge for the use of public data are permissible; societies can choose how much free use they want to allow.

While the regulation of data use, machine learning and AI has been identified as an important research field by legal and privacy scholars (e.g., Fairfield 2017; Lane et al. 2014), normative

---

[2] https://support.google.com/mapcontentpartners/answer/9359673?hl=en&ref_topic=21600 [accessed 30 July 2019].

political theory is lagging behind. In particular, theorists should focus on the inequality created between those at the center and those at the periphery of technological development. Some firms producing AI systems have or are about to achieve enormous technological, economic, and social influence. At the same time, AI promises to be a disruptive technology that will displace, not only enhance many jobs (Brynjolfsson and McAfee 2016).[3] While the winners may get very wealthy, a greater number of people will likely be losers in this transition. And while there may be a prospect that new jobs are created and an overall gain in welfare is obtained in the long term, this is of little comfort if one's life lies in the short term (Frey 2019). Another sign of things to come is the inequality created between the employees of successful internet companies who have experienced fast-rising salaries and the employees in more 'bread-and-butter' firms with traditional business models, who see their salaries stagnate (Bloom 2017).[4]

In this paper, instead of drawing on one's preferred theory of egalitarian justice, condemning the unequal distributive outcome of AI in isolation (e.g., Rawls 1999), we propose to train the focus on the production *inputs* and how their use causes inequality. More specifically, we ask whether the large-scale use of data for the creation of proprietary artificial intelligence systems provides grounds for requiring payment from the owners of artificial intelligence. This approach also helps to shed light on the nature of a special public good: public information. We

---

[3] Acemoglu and Restrepo 2018 provide a formal model to explain how this can have a negative impact on demand for labor.

[4] The dawn of countless algorithms to classify, assess and sort clients, applicants, employees, etc. also creates a power differential between designers and takers of such regulatory systems, and often a lack of accountability, too (O'Neil 2016).

will argue that limiting access to or charging for the use of that public good is, given certain conditions, permissible and often desirable. Developing rules for fair use of public data is most urgent now because 'datafication', data scraping and automated learning technologies, change the way data are created and used.

The next section of the paper introduces the notion of the 'global information commons', a public good consisting of openly accessible information. We then explain how artificial intelligence systems learn and why learning typically requires large amounts of information. The global information commons are a special public good because they are not undersupplied. We proceed to present the Free Use Argument for the global information commons. To show that it is unsound, we provide two different arguments against it. The Externality Argument points to the inequality that arises as a by-product of free information use. The Use Expectation Argument explores whether information providers have a right to control the use of information they add to the commons. The last two sections sketch some policy proposals and conclude.

THE GLOBAL INFORMATION COMMONS

The set of publicly available information on the internet, in libraries, and other open repositories is sometimes referred to as a global commons (Stiglitz 1999, 315; Hess and Ostrom 2003) or the creative commons (Creative Commons 2018; Boyle 2008, ch. 8). We will use the term *global information commons* to refer to all information that is available without access restriction. The internet is a major factor for the expansion of these commons; a largely decentralized system of contributors has created a rich source of information. Much of this information is provided by unpaid enthusiasts. The global information commons include blog posts, websites, shared videos, public social media posts, open access archives, and much more.

Not everything that is accessible to everyone may be used legally in every way possible. For example, the licensing terms of Wikipedia contain a 'share alike' clause, which states that '[i]f you alter, transform, or build upon this work, you may distribute the resulting work only under the same or a compatible license.' (Wikipedia 2018). This is a non-trivial use restriction. By contrast, information that is in the 'public domain' can legally be used in any way imaginable without any obligations to the creator (Boyle, 38-9). For instance, Beethoven's music scores are now in the public domain because intellectual property rights have expired. And the CIA Factbook is in the public domain because the US federal government waived any rights to it (CIA 2018). The interesting question is: what kind of use restrictions ought to apply to the global information commons? An answer requires a better understanding of what the information commons are.

The information commons are characterized by four properties:

(1) The information available is valuable to at least some users.

(2) The information is non-rivalrous in consumption, so that, for the existing users, the marginal loss in this value from adding an additional user is zero.

(3) Excluding users is difficult.

(4) Most information is supplied on a voluntary basis.

The first property tells us that the information commons are a good.[5] Properties (2) and (3) are what most economists take as the two necessary features of public goods (Samuelson, 1954;

[5] The global information commons are a good in many ways: as an educational resource, as a repository of human knowledge, a resource to spur technological innovation, etc.

for an application to data see Brynjolfsson and McAfee 2016, 62; Pollock 2009). First, information is, by its very nature, non-rivalrous in use. When a good is rivalrous, one person's use diminishes the good for others. Information is non-rivalrous because an unlimited number of users does not diminish the good in any way.[6] Second, once the information is available on the internet, it becomes hard to contain the circle of users.[7] This suggests that the information commons are a public good in the traditional sense. (4) marks out a peculiar feature of the information commons: unlike many public goods that are only supplied with adequate incentives, the information commons are to a large extent created without any need for incentives (Brynjolffson and McAfee 2016, 64).[8] This fact will feature prominently in the Free Use Argument we criticize below.

---

[6] It is often the case that the market value of information goes down the more people know about it. Therefore, even if the use of others does not diminish use value, it can diminish market value. That may constitute an argument to support intellectual property rights, as Cwik 2016 points out. But this is not the issue at stake in the present argument. The problem is not that the subjects creating data want to keep exclusive commercial use rights. The problem is that others make use of data in a way that they object to, even if they consented to the data being part of the global information commons.

[7] Encryption techniques and digital rights management techniques offer effective use restrictions, but once the information is publicly available somewhere, exclusion becomes hard, if not impossible.

[8] Some other parts are a by-product of the desire to self-promote, seek attention, or to pursue business interests, but even in those cases the information is voluntarily provided.

Since we focus on the global information commons, other challenges related to 'big data' are beyond the scope of this paper, especially those big data sets that are not part of the global information commons. Since data[9] is valuable, major tech companies also pursue business models that entice their clients to contribute ever more data over which they have exclusive control, outside of the global information commons. While this is an important topic in its own right (Lanier 2014; Mayer-Schoenberger and Ramge 2018, ch. 8; Sadowski 2019), this paper focuses on the global information commons. We do this because we think that there are already good normative reasons for limiting walled, non-public systems in their data harvesting habits, for either principled (Kitchin 2017, ch. 3; Lessig 2004; Solove 2012) or pragmatic reasons (Mayer-Schoenberger and Ramge 2018, ch. 8). Our point in this paper is that, even if the walls are broken down and more data becomes public again, important normative issues about the use of the global information commons remain.

MACHINE LEARNING: PRINCIPLES AND ECONOMICS

There are three main kinds of artificial intelligence, one that exists now and two that are predicted to come into existence in the future. These can be distinguished as 'artificial narrow intelligence' (ANI), 'artificial general intelligence' (AGI), and 'artificial superintelligence'

[9] We use 'data' and 'information' interchangeably. Strictly speaking, information is data that has been processed, organized and analyzed. Information is therefore of greater value for the goal of acquiring knowledge. For this paper, the distinction between information and data is less important, however.

(ASI). ANI specializes in one task; AGI is generalized across a wide range of tasks and approximates human-level intelligence; and ASI outpaces human intelligence exponentially (Urban 2016). Each of these raise intriguing moral and political questions, but we're going to set ASI aside in this paper and focus on ANI and AGI.

Examples of ANI include the chess-playing technology 'DeepBlue', various map and navigation apps, and the Go-playing technology 'AlphaGo,' which beat the human grandmaster Lee Sedol at Go in 2016 (Metz 2016). What many of these systems have in common is that they rely on publicly available information, which has been and continuously is created and made available by individual contributors across the world. Thus, for example, AlphaGo learned Go strategy through extensive training, based on records of both human and computer players. Its initial training was based on historical games, from a database containing some 30 million individual moves (Metz 2016). It then trained by playing against counterparts of itself (Silver & Hassabis 2016). Another example is automated language translation algorithms, learning from corpuses of translated texts, such as European Union documents (Koehn 2010).

Not all ANIs depend on such data-led training. Strikingly, Google's DeepMind team created an algorithm that can learn how to play classic Atari computer games like 'Breakout'. In this case, learning depends entirely on repeated attempts and not on data from human players (Mnih et al., 2015). More recently, the DeepMind team even presented a Go learning algorithm that masters the game by just playing the game a great many times against itself (Silver et al., 2017), without any human moves as input. This shows that not all artificial intelligence depends on prior training data. Sometimes 'brute force' repeated attempts, combined with reinforcement learning, do the trick. However, Atari games and Go are particularly well-suited for this form of learning because the basic rules are very well-defined, the games are deterministic, and the algorithm gets the chance to train by playing the game as often as necessary.

Most tasks for artificial intelligence will not be learnable in such a controlled environment. Steering a self-driving car, for example, requires navigating a complex and changing environment and dealing with random events. And, crucially, meaningful training depends on input from the real world, not only simulations. Therefore, learning from (many) mistakes is simply not an option for self-driving cars. Even seemingly more closely circumscribed tasks, such as translating from German to English, cannot be mastered by AI without drawing on extensive training data. Without data, the system simply would not get enough feedback to eliminate mistakes.

The upshot is then, that in most cases, and certainly for more complex tasks, ANIs are parasitic on constantly expanding pools of information, and AGI's are likely to be so to an even greater degree because they need to learn from a diverse set of situations to deal with complex environments (Tegmark 2018, ch. 4). The *functionality* of most ANIs is made possible by the information people provide. The important point here is that the most interesting and promising ANIs are not only capable of *learning*, in many cases their functionality *necessarily* depends on training data.

The most important machine learning approach these days is the use of artificial neural networks, taking cues from biological brains. These networks have led to major learning breakthroughs. Image recognition, for example, is an area for which the combination of artificial neural networks and extensive training with real-world data is key to progress. For instance, trained artificial neural networks now match dermatologists in predicting the malignancy of skin lesions (Esteva et al., 2017). Interestingly, prior to be trained with skin lesion images, the AI learned how to recognize patterns by drawing on publicly available images. Such 'unsupervised learning' picks up on structures in the world without needing feedback from labelled data and opens new avenues for the training of AI (LeCun 2015).

It is not a coincidence that the major players in the field of AI are keen to collect as much data as they can. To some extent this data is useful to understand their own customers better, or to provide their users with the data they need. But another reason is that the development of AI cries out for additional data. For conventional information gathering purposes such as statistics, data quickly loses marginal value after a certain amount is accumulated (i.e. when there is enough data to make a statistically valid inference). But for machine learning purposes, more data does not lose marginal value as it allows AI systems to master more complex tasks (i.e. make compound inferences).[10] When training machines to perform complex tasks, only a lot of data helps, and a lot of data helps a lot:

> '… the returns to data may decline only gradually or there may even be increasing returns to data if more sophisticated tasks are disproportionately more valuable. This is consistent with the empirically-observed dominance of the data economy by a few large firms.' (Arrieta-Ibarra et al. 2018, 40)

There are at least three reasons why the commercial use of machine learning is likely to lead to a few very big and very important companies:

1. Training AI systems able to interact with complex environments requires enormous data collection, storage and processing capacity, creating significant barriers to entry[11];

---

[10] See Posner & Weyl 2018, 224-230, for an excellent analysis of how data sometimes has increasing, rather than decreasing marginal returns for the purpose of machine learning.

[11] For a small but instructive example, consider the enormous amounts of 3D mapping data needed to navigate an autonomous car and other mobile robotic devices through a city like London. See Hook 2018.

2. The market for artificial intelligence systems tends to have a winner-takes-all structure. The best-performing systems are likely to become the preferred choice for all consumers (Brynjolfsson and McAfee 2016, ch. 10);
3. Many tech companies have created services that ensure a constant stream of new data, which helps to improve their AI systems further, a positive enforcement loop.[12]

Each effect on its own is sufficient for undermining market competition. But taken together, the effects are likely mutually reinforcing and lead to the emergence of a small number of major players, unrivalled in economic power and influence, not unlike the current market for major internet and tech companies, though possibly more extreme. One can already observe that the market-leading firms in digitalization experience fast growth, while others lag behind. On the level of employees, a 'hollowing out the middle-skill portion' of the job market is predicted, leading to a widening income gap (McKinsey 2015, 12). While one cannot know for certain, we will work with this empirical assumption: without regulatory intervention, and provided that free training data continues to be in plentiful supply, the market for AI systems will likely contribute to substantially increased economic inequality. Furthermore, the greater dominance of a few major corporations will make it increasingly difficult for others to challenge their market position, entrenching the rising economic inequality. It will also add to inequalities in technological dominance of private actors and likely an increased social and political influence of the future owners of the most advanced AI systems.

[12] See Weber 2017, for examples of this positive feedback loop on an international level.

USE, NOT PROVISION, IS THE PROBLEM

Since the information commons is a public good, one might think that it faces the same problem as most public goods: free-riding on the freely provided public good. Upon closer inspection, however, free-riding in provision is not a concern that pertains to the information commons. The real problem, we will argue, is unrestricted use.

Consider, for contrast, a classic example of a public good: a light house signal. Using the signals to navigate is non-rivalrous because use by additional ships does not diminish the use value for others. Exclusion of users is difficult because ships operating in open waters are hard to monitor and charges consequently difficult to enforce. Such a public good is costly in production. In a free market it will be under-supplied (or not supplied at all) because users have an incentive to free-ride on others' provision and suppliers will not be paid for the socially optimal level of service.

This example helps us to see how different the information commons are in that regard. For a start, accessing the good without contributing to its provision (which would otherwise be free-riding) is not a problem; in fact, it is invited. Providers of information want other users to access the information – that is the very point of posting information publicly. And notably, the information commons do not suffer from information under-supply. Instead, what we see is a remarkable level of voluntary information provision by a great many members of the public. Some of this apparent voluntariness may be driven by hidden non-altruistic motives. Still, the fact remains that the global information commons have been filled with content in a largely non-enforced, non-regulated fashion and with only limited monetary incentives. This shows that the information commons are very much unlike lighthouses: they get provided in good measure without enforcement.

The relevant question is therefore *not* how we can ensure that everybody contributes. Free-riding in provision, the standard problem for most public goods, is not the concern. The real issue is about the permissibility of using the information commons in excessive, unanticipated ways.

THE FREE USE ARGUMENT

One may think that charging for public information is always impermissible because information is the perfect public good. This argument is often appealed to, but rarely stated precisely.[13] Justice Brandeis alludes to it here:

> 'The general rule of law is that the noblest of human productions – knowledge, truths ascertained, conceptions, and ideas – became, after voluntary communication to others, free as the air to common use.'
>
> (International News Serv. vs. Associated Press, 250, Brandeis dissenting; see also Benkler 1999)

Brandeis appeals to the idea that information, once it has been passed on to others, becomes part of a commons that is free to use for everyone.

Not all uses are allowed, of course: works that are protected by copyright may not (legally) be reproduced or re-used in specific ways without permission, but if they are offered to the public they may still be read, digested, and processed. Similarly, information protected by patents is

[13] A notable exception is Benkler 2006, chapter 2. The argument has also been suggested to us by colleagues in discussion.

legally restricted in specific uses, though may still be freely consumed in other ways. Copyright and patent restrictions are in place to prevent under-supply of such content. However, for the sake of this argument we can put prohibited uses of copyrighted or patented material to one side; of interest to us are *non-appropriating uses* such as reading, digesting, and processing the information. The term 'use' is to be understood in that sense in the argument to follow.

In our reconstruction[14], the argument from free use runs as follows:

The Free Use Argument

(1) The information commons are a perfect public good and thus non-rivalrous in use.

(2) The information needed for the information commons is not under-supplied.

(3) If a good is non-rivalrous in use and not under-supplied, no one is adversely affected by the use of the good.

(4) Because the information commons are non-rivalrous in use and not under-supplied, no one is adversely affected by their use.

(5) Using a public good free of charge without adversely affecting anyone is always permissible.

(6) Therefore, using the global information commons free of charge (even extensively and for private gain) is always permissible.

---

[14] A related argument against intellectual property protection is analysed and dismissed by Himma 2005 and Moore 2012. Cwik 2016 provides a very helpful critical review of that debate. Our focus, however, is on data, not intellectual property, and our goal is different: to show that some uses of the information commons may be restricted.

An important upshot of the Free Use Argument is that, if sound, it rules out any restrictions on or charges for the use of the global information commons, rendering such restrictions impermissible. We aim to show that the Free Use Argument is unsound by demonstrating that premises (3) and (5) are false. If the free use argument is not sound, we have removed an important argument against regulating the use of the information commons, weakening the case against regulation. If and how a society should attempt such a regulation, of course, depends on a number of normative and practical factors that need to be balanced. We will make some proposals in that regard at the end of this article.

The proponents of the Free Use Argument might insist that information is a special public good. While tangible goods can be owned, information cannot be owned in the same way, the argument goes. Sure enough, there are intellectual property rights, but even the most stringent such rights do not prevent users from consuming and digesting information. This would lead to the more restricted:

(5*) Using public information free of charge without adversely affecting anyone is always permissible.

We take (5*) to be the most promising rendering of the premise. It usefully restricts the Free Use Argument to public information. And, in combination with (4), it rules out problems arising from under-provision, which would have made usage fees permissible. In case of the global information commons, however, under-provision is simply not an issue. Therefore, the argument goes, the only reason for use charges is eliminated and freedom of use is the logical outcome.

However, this turns out to be too hasty. Our strategy against the Free Use Argument is based on the idea that people creating public goods like the global information commons may be permitted to regulate the use of the public good if this is in the public interest. To determine

whether use ought to be regulated, two questions should be considered. First, is the free use of the public good serving publicly endorsed goals, or does it create unwanted negative externalities that regulation could mitigate? Second, is the way the public good is used in line with the intentions that its creators had in mind (or reasonably should have had in mind) when creating the public good?

To pursue these two questions and defeat the Free Use Argument, we first train our attention on premise (3), which we aim to reject in the next section. We revisit Premise (5*) in the section thereafter. However, it is worth pointing out that positive answers to our two questions only provide pro tanto reasons in favor of regulating the global information commons. Whether we ought to regulate also depends on whether regulation would be overall beneficial, addresses the concerns raised by our questions, and does so in an effective way.

THE EXTERNALITY ARGUMENT

The Externality Argument shows that the empirical claim of premise (3) is often false, and indeed false in the case of how major AI companies use the global information commons. Demonstrating this is theoretically quite straightforward: we need to show that, even though the information commons are non-rivalrous in use and not under-supplied, there are adverse effects on others caused by some forms of use. If we succeed in doing so, then (4) is false, providing a counter-example to (3).

The information commons are supplied on a sufficient level by volunteers. What could be wrong with using this pool of information for free? Proponents of the free use argument suggest that no one is adversely affected by the intensive use of a good, as long as the good is non-rivalrous in use and the provision problem has been taken care of. Admittedly, in a spe-

cific narrow sense that is true: using more of the information commons does not take the information commons away from anyone. The resource does not diminish, and the other users are not affected *in their use of the information commons*.

But why consider only direct effects? If you are concerned about the distributive impact of providing free use, then you will worry that other users are affected in an indirect way: by increasing inequality if the producers of highly capable AI will become very rich and influential. Economic inequality is the most immediate effect, but this is likely to be followed by inequality in social influence, unequal access to information, and political inequality.[15] Consequently, allowing such very intense use is far from harmless: while it does not diminish the resource, it has side effects that are better avoided: the inequality created as a negative externality of extensive use.

The structure of this argument is familiar to political theorists. There are economic transactions that look unobjectionable if one only considers the transactions on their own terms, but objectionable when taking into account the distributive effect. For a famous example, take Robert Nozick's entitlement theory, made vivid by the 'Wilt Chamberlain' example (Nozick 1974, 160-4). Nozick envisages a set of voluntary transactions among property owners or individuals exercising their self-ownership. Specifically, many basketball fans are prepared to pay extra for seeing basketball star Wilt Chamberlain play, making each fan only a little poorer but Chamberlain a lot richer, leading to economic inequality. Nozick, famously, sees nothing

[15] Sadowski 2019 argues that data can be understood as a form of *capital*; not only valuable in itself, but also a means to *generate* value, monetary and otherwise.

in the transactions that can be objected to and consequently denies that any measures to interfere with these voluntary transactions could be just. Because the transactions are voluntary, no one involved can object, Nozick maintains.

Many critics were quick to point out that, even if all transactions are unobjectionable according to the criteria we use to assess such transactions by themselves, the aggregate result may still be objectionable (Cohen 1995, chapter 1; Dworkin 2000, 110-12). One can criticize the outcome along two different lines. First, the *distribution* arising from many individual transactions may be objectionable even if each particular transaction is not – and, indeed, the individuals involved in the Chamberlain example might not have agreed to pay if they had known of this distributive outcome (Cohen 1995, 23). It does not follow, as Cohen points out, from people being willing to pay to watch Wilt play that they are willing to pay Wilt to play (Cohen 1995, 26). Second, the Chamberlain fans are faced with the choice of either watching Chamberlain and contributing to his riches, or not watching Chamberlain to avoid contributing to his riches. What they are missing is a third option: watching Chamberlain play without making him rich. It is this third option that they arguably should be provided with – and the one egalitarian, basketball-loving citizens would prefer. These two considerations support redistribution or taxing the transactions. For example, there might be a case for putting a sales tax on Wilt Chamberlain's tickets and redistributing the income.

The structural similarity to the use of the information commons should be apparent. Replace the transactions in the market with the contribution to and use of the free information commons. In both cases, the actions considered on their own merit are not the problem. The contributions to the commons are voluntary, and the use of the commons is prima facie permitted because the commons offer a non-rivalrous public good in sufficient supply. The problem with the free use of the commons is the distributive effect: The availability of so much free information creates a social possibility that some companies may exploit; and if they do, this creates

more inequality than many societies are (and ought to be) prepared to accept. And these effects are further exacerbated by the monopoly-like status that several such companies enjoy, which undermines competition and the viability of customers opting out from their data-collecting services (Posner & Weyl 2018, 230-239).

The contributors to the global information commons are effectively forced into a choice between two options: they can either continue to supply information in the usual way, thereby accepting that producers of AI systems will use the information for free to become rich and powerful. Or they can withdraw from the internet, stop contributing to the global information commons, to prevent the extreme inequality caused when 'subsidizing' AI companies with free information. This is an unenviable choice, as the opportunity costs for withdrawing from the internet are substantial and increasing.

A third option ought to be on the choice menu: contributing to the global information commons *without* turning AI producers into new super-charged Wilt Chamberlains. Since this option is currently unavailable, it is not true that the free use of the global information commons by AI producers is without adverse effects. The adverse effect is the inequality it produces. And just like basketball fans do not make a fully voluntary choice when selecting between making Chamberlain rich or stopping attending games, contributors to the information commons do not make a fully voluntary choice when selecting between making AI producers rich or stopping contributing to the global information commons.

There is, however, a relevant difference between Nozick's Chamberlain and our voluntary contributions to the global information commons: the individuals deciding to pay to see Chamberlain play decide to part with their money in exchange for the performance and they know that that money will end in Chamberlain's pockets, eventually making him rich. And even if they have not consented to the pernicious effects of the resulting inequality, they have – in

some minimal sense – consented to these transactions and their predicable aggregate effect. By contrast, the voluntary contributors to the global information commons do not engage in a market exchange. Their mode of interaction is not intended as payment for a service provided. Moreover, when Nozick's basketball fans pay to watch Wilt play, they are offering one small contribution to a much larger effect *of the same type*. When people contribute content to the information commons, however, they produce an entirely unrelated aggregate effect. Therefore, since the inequality that arises from the training of artificial intelligence systems is a byproduct, an unintended external consequence, to an even greater extent, our externality argument is even stronger than Cohen's argument against Nozick. The unexpected nature of use and effects also links the Externality Argument with the Use Expectation Argument, to which we will turn in the next section.

The upshot of this discussion is that premise (4) is typically false. The unlimited use of the information commons causes negative externalities in the form of inequality. Since this is an adverse effect, a counter-example to premise (3) has been provided, and therefore that premise must be rejected. It remains an empirical matter, of course, how much inequality is caused by letting AI builders use the whole information commons for free, but the effect is likely large.

It is possible to resist the externality argument by denying the egalitarian starting point. Nozick, famously, took such a libertarian view, rejecting the idea that equality is important enough to justify interfering with self-ownership (Nozick 1974, 160-74). In a similar vein, one might suggest the inequality externality does not matter enough to undermine the conclusion of the Free Use Argument. This position, however, flies in the face of even minimally egalitarian theories. And while Nozick could argue that taxing individuals is a significant interference with self-ownership (he famously compares taxes with slavery), taxing data only affects self-ownership in the most minimal way: it prevents the producers of data from giving away their

data to tech companies for free. But most contributors never intended to do that anyhow, as we explain in the next section.

THE USE EXPECTATION ARGUMENT

While we have just seen that the empirical part of premise (3) is often false, there are also reasons to reject premise (5*). Recall, the premise at stake is:

(5*) Using public information free of charge without adversely affecting anyone is always permissible.

On the face of it, there is a strong case in favor of completely free information, without any use restrictions. Many internet activists embrace a culture of free and open sharing, with content that is free for everyone to use, distribute, and remix. If intellectual property legislation and copyright laws become too tilted in favor of protecting vested commercial interests, so the argument goes, the initiative of creators is constrained (Lessig 2004, 29).[16] The fight against excessive copyright restrictions appears to have blindsided internet activists in one respect, however: the danger of unwittingly supporting the commercial interests of enterprises that consume large amounts of data.[17]

---

[16] It is worth noting that the popularity of this view might also be due to path dependency and how the internet has historically developed, rather than independently formed user intentions (see e.g. Posner & Weyl, 209-213). This, of course, further strengthens our argument.

[17] Some, however, have picked up on this danger, most notably Jaron Lanier (2014).

In brief, our argument against (5*) is this: First, taking information from others is normally impermissible. Second, it is permissible with consent, but in the context of the global information commons the consent is implicit. Third, whether one may assume implicit consent depends on what is at stake and how far the actions depart from normal expectations. Fourth, using data to train AI involves high stakes and is a significant departure from expectations. Therefore, fifth, only a strong form of consent suffices. Finally, since such a form of consent has not been given, using the global commons on a large scale to train AI is presumed to be impermissible, unless there are other overriding reasons. We will now state this argument in detail.

Let us begin with the observation that the taking of information is normally only permissible with consent. For example, it is impermissible for someone to secretly copy your research notes, regardless of whether they use them to benefit society, and regardless of the fact that you do not lose your notes if they are copied. By contrast copying them *with* your consent is permissible. Therefore, the real question is what the contributors to the global information commons have consented to.

In the case of the global information commons, most consent is implicit – it lies in the act of contribution and publication. Internet users post information for different reasons. They may want to help others to find information. They may want to speak to an audience or advertise their work or their thoughts. But very few users expected that they will help train AI systems that will make their owners become very powerful und rich. Have they consented in implicit fashion nevertheless?

On a first pass, the answer seems to be 'yes'. For example, a DIY expert may create a website explaining how to fix plumbing issues with the expectation of helping users save on repairs. If a user learns the necessary skills from that website and sets up a lucrative plumbing business,

then it is implausible for the website creator to not have expected this and deem it impermissible. By making content available on the internet one implicitly consents to such use scenarios. The contributors to the information commons know this (or ought to have known, anyway).

What kind of uses can contributors reasonably expect? The internet changes quickly, so it is not easy to give a general answer. However, one structural factor about use remained stable until recently: the use of content was constrained by the capacity of an agent being able to *understand* and *interpret* content. Understanding and interpretation are capacities unique to humans. For this reason, a contributor to the global information commons was – until very recently – assured of one fact: a human would be needed to interpret and understand their contribution. And since human capacities for processing content are limited, no single individual would be able to process excessively large parts of the information commons on their own. The natural limitation of the human ability to take up content created an *individual use expectation* for the global information commons.

The individual use expectation has been shattered with the advent of automated learning and AI. A new class of users is arriving on the scene: learning AI system with their insatiable hunger for new data and the ability to process information very quickly. This, then, is the use that the contributors to the information commons have (typically) not expected and (almost never) intended.

What kind of consent is required under such conditions? Tom Dougherty's (2015) work helps clarify this. Following Dougherty, the higher the stakes involved in a consensual transaction, and the more the relevant act runs counter to normal expectations, the more important it is to obtain consent grounded in common belief among the affected parties. This ensures that everyone has the same beliefs, and everybody knows about these beliefs. Common belief guarantees that all parties are 'on the same page', ruling out misunderstandings about what was

agreed. Clearly, most tech companies harvesting the global information commons have not obtained consent in the form of common beliefs. They have, at most, obtained an implicit consent for some uses, stemming from the contributors' decisions to post publicly.[18] But this does not suffice as consent for AI training use because the stakes are especially weighty and the new use unexpected: the amount of data used as well as the effects of that use are as important as they are surprising, as we have seen. Under such circumstances, only consent with common belief will do.

If use goes beyond normal expectations or norms, the onus lies on the data collector to show that the subject would have good reasons to consent, or that the societal benefits are so overwhelming that one may proceed without consent (Barocas and Nissenbaum 2014; Nissenbaum 2009, ch. 7). The first condition is unlikely to be met, the second would require a careful assessment and societal debate that is yet to take place. In addition, there is a systematic mismatch between the intentions with which users provide data (such as knowledge gathering and social networking) and the purpose for which corporations need the data (i.e. to target advertising and train their AI systems).[19] There is also a related mismatch between the often idealistic motives of the contributors and the purely commercial interests of corporations using the data.

---

[18] Indeed, Sadowski 2019 notes that end-user license agreements (EULAs), commonly used to obtain consent from contributors 'are long, dense legal documents […] designed to prevent even the most enterprising person from being informed of the binding terms and conditions' (7). Because of this, he argues, the type of consent acquired by data corporations through EULAs 'bears little resemblance to common meanings of consent' (8).

[19] See Posner & Weyl 2018, 220-224.

Without consent, premise (5*) is not plausible, despite the 'free information' rhetoric. Even if no one is adversely affected by the use of public information, the authors or contributors of that information have a case against such use because the right type of consent is missing. Proponents of free use would owe us an overriding reason important enough to waive the consent requirement, a reason difficult to provide.

One might object that contributors to joint voluntary projects often have all kinds of expectations, but these expectations do not necessarily give them the right to control the use of their product. For instance, if Ann helps to plant a local community garden on public land, Ann cannot expect to have control over who uses the garden and how. If, for example, Ann has the intention to create this garden as a resting place for the elderly, her expectation does not give Ann the right to chase away the young office worker. However, one might retort that Ann's use expectation was unreasonable from the start. By contrast, the contributors to the global information commons had a reasonable expectation, one that is undermined if the information is automatically read, collated and processed by machines to train AI systems.

A second line of objection points to the fact that nearly all intellectual activities are built on those of others. Indeed, one powerful argument in favor of free use is the innovative potential of the internet, where information is often given new and unexpected form, improving and enriching the commons. Some might argue this function is the very reason why the information commons are a 'good'. Does the use expectation argument lead to the absurd conclusion that all such use is impermissible unless consent can be shown? For example, if a novelist writes a bestseller, she benefits from the input of many others who have shaped the language she is using. And these contributors have not consented to the specific use. Nevertheless, it is

implausible that the novelist's use of language is impermissible.[20] For a response we need to explain the difference between the language and the information commons case. First, we maintain that language contributors are implicitly consenting. Language is a social product, and it has always been used for diverse purposes. Writing a commercially successful novel is well within the range of expected use, so implicit consent can be assumed. By contrast, the use of data to train AI is a novel, unexpected kind of use. We also observe, second, that intuitions about the language argument flip when pushed to extremes. Imagine a clever businessman who unexpectedly comes into contact with extra-terrestrial aliens and sells them a complete inventory of the English language, making a fortune. In that case, the people contributing to English as a public good can complain that they did not consent, neither explicitly nor implicitly, to the sale and its distributive effects. This fictive example shows that the difference lies in the novelty and scale of the new use, in line with our argument against public data use to train AI. The more novel or unexpected the use, and the larger the stakes, the more important it is to obtain consent as common belief.

A final objection denies that consent is needed for using data once it has been added to the global information commons, as contributors relinquish any moral rights to that data with the act of publication. However, this misconceives the complexity of rights given up, retained and created when contributing to a public good.[21] For a simple example, if you contribute to a community potluck, you are happy for your cake to be eaten by neighbors (to which you have

---

[20] We thank Jacob T. Levy, the editor, for raising this objection.

[21] A helpful framework to think about this complexity is Helen Nissenbaum's (2009, ch. 7) contextual analysis of informational norms, especially her discussion of 'transmission principles' on p. 145ff.

consented by contributing) but you can rightfully object to it being sold for someone else's profit (to which you have not consented). The example shows that contributors to public goods normally retain the right that their contribution is used in line with consent (often implicit) about the purpose of the public good. Contribution might also create a new right to jointly regulate the public good created.

INSTITUTIONAL PROPOSALS

Contributions to the global information commons tend to have the unwanted side-effect of promoting inequality. This is for two reasons. First, we provide the developers of AI with a crucial resource for free, as explained above. Second, and in addition to the reasons discussed above, there are structural reasons why big-data-driven economies tend to benefit large companies: increasing returns when scaling up (Arthur 1989), network effects (Zhang et al. 2015), and feedback effects due to improved predictive power of larger datasets (Mayer-Schoenberger and Ramge 2018, ch. 8). These mechanisms give advantages to those who gather most data and those who enter the market first, reinforcing the inegalitarian trend.

We now want to sketch two proposals to tackle this problem: creating a new market for data labor, and asking heavy users to pay for the services of the global information commons by introducing a progressive tax for data use.

Lanier (2014, 16), Arrieta-Ibarra et al. (2018), and Posner & Weyl (2018, chap. 5) propose a transition towards a 'Data as Labor' market structure. The core idea is to make data the property of its creator, enabling the creator to charge for data use. In terms of the Free Use Argument, the proposal works against premise (1) by changing the property regime from commons to private property. Contributors to the global information commons are turned into owners of data, and this data can only be used if the owners agree.

Privatizing the global information commons in that way will come with some problems, however. For a start, the system can only be implemented if it is possible to process very small payments efficiently. If the transaction costs dwarf the economic value of the separate pieces of data, the market will dry up. In addition, the buyers of data also need to check whether the data they buy is new and useful, not recycled old data or meaningless noise. Checking this will surely be automated, but the computational power needed to do so might well be so expensive that each small, single transaction becomes pointless. A data as labor payment system also requires a method for assessing the value of each item of data and information. This is a difficult task, given that the value of data is typically only realized in aggregation (as 'big data'), and often in ways that were not anticipated initially.

Perhaps these implementation problems can be overcome. But there are other practical and normative shortcomings. First, privatizing the commons might lead to an under-supply or deterioration of the global information commons. If contributors can charge for their information while access to the global commons stays free, then an additional incentive is created for gathering data in closed networks rather than in an open access network. This is because the data is relatively more valuable if one has exclusive access to it. The market price for exclusive data will be higher, so that contribution to the Global Information Commons are financially less attractive and the commons are likely to deteriorate.

The second, normative problem with the 'Data as Labor' proposal concerns distributive effects. With payments flowing from data users to data owners, the resulting distribution might be more equal. But it might also reward the wrong people. Content producers are often highly educated and economically advantaged (Hargittai and Walejko 2008; Schradie 2012). Additional payment for their services might lead to additional benefits for the already better-off. Then again, 'Data as Labor' might also lead to a new class of self-employed data laborers with precarious incomes, the 'digital proletariat' (Economist 2018). Whatever the outcome might

be, 'Data as Labor' does not directly tackle the externality problem because it leaves the distribution to a new market for data labor, with difficult to predict distributive implications. However, 'Data as Labor' would go some length to address concerns arising from use expectation, as the data producers literally take ownership of their data and can ensure that their use expectations are respected.

A second policy option is to think about the global information commons not as a set of many small pieces of individual property, but as a commons that can be governed as a whole (Morozov 2018). One way to govern these commons is to charge for extensive use. In practice, this will amount to a tax on data used.[22] The actual tax base could be the amount of data accessed. Alternatively, one could use imperfect proxies, such as computing power or data storage capacity, which might be easier to determine by measuring physical infrastructure. A well-designed 'data tax' would tax the most intense, industrial uses of the global information commons, without creating disincentives for smaller users of the global information commons. It should be a progressive data tax. It would allow free or almost free access for human users to promote unhindered exchange of ideas on the internet. Ideally, it would target those corporations that profit most from data use.[23] To reduce the negative externality of inequality, the

[22] See Chris Hughes (2018) for a proposal and some design challenges. For a related debate on a tax on robots, see, e.g., Englisch (2018) with further references.

[23] We thank Laura Valentini for suggesting this to us. One theoretically attractive option would be to make the corporate tax rate dependent on data use. In practice, however, this will be difficult to implement, as it creates incentive to choose a corporate structure that separates data use from profit.

tax revenue could be used to widen access to and participation in the global information commons.

A progressive data tax will not give the individual providers of information the right to charge for their individual contributions. However, the tax targets, more directly than the other policy proposals, the central problem we identified: that free access for all produces unacceptable levels of inequality. The externality argument can be used to justify such a tax. If implemented well, it could reduce the inegalitarian effect of free data use, and the tax revenues can be employed to tackle inequality directly. For these reasons, we think that a progressive data tax is the approach that merits most serious attention, though more work is needed to develop a concrete proposal.

CONCLUSION

The internet, once imagined as a sphere of free information exchange among equals, has turned into a resource. Intended as a commons to be useful for all, it is now the well of data for private enterprises, used by AI companies to train their self-learning systems. These companies are currently allowed to use the resource without charges and without restrictions. One may think that this is only right: since using information does not diminish the use value for others, no one is negatively affected. But this is an erroneous thought: the free big data supply increases economic and other inequality, and it disrespects the use expectations of many content providers. Free use does have adverse effects, and this is why the use of the information commons may and ought to be regulated.

The global information commons are not owned by anyone. But that does not mean that their use should be a free-for-all to the benefit of the biggest and most powerful players. A progressive data tax could be an effective policy instrument to prevent the inegalitarian trend. Other

policy options might be available. How and to what extent we regulate the global information commons is a political question that requires urgent public attention.

Acknowledgements: We are grateful for extensive comments from Laura Valentini, Alex Voorhoeve, Mathias Koenig-Archibugi, Tim Meijers, and Annette Zimmermann. We would like to thank audiences at the University of Aarhus, the LSE Department of Government Colloquium, the Luc Bovens workshop, the LSE Political Theory Graduate Conference, the University of Groningen (especially Frank Hindriks, Andreas Schmidt and Ryan Doody), as well as a great number of our students. Special thanks to the anonymous reviewers for this journal, who provided detailed and constructive feedback.

REFERENCES

Arrieta-Ibarra, Imanol, Leonard Goff, Diego Jiménez Hernández, Jaron Lanier, and E. Glen Weyl. 2018. 'Should We Treat Data as Labor? Moving Beyond "Free."' *American Economic Association Papers & Proceedings* 108: 38-42.

Arthur, W. B. 1889. 'Competing Technologies, Increasing Returns, and Lock-In by Historical Events.' *The Economic Journal* 99: 116-131.

Barocas, Solon, and Helen Nissenbaum. 2014. 'Big Data's End Run around Anonymity and Consent'. In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Fay Nissenbaum, 44–75. New York: Cambridge University Press.

Benkler, Yochai. 1999. 'Free as the Air to Common Use: First Amendment Constraints on Enclosure of the Public Domain.' *New York University Law Review* 74: 354-446.

Benkler, Yochai. 2006. *The wealth of networks: how social production transforms markets and freedom*. New Haven and London: Yale University Press.

Bloom, Nicholas. 2017. 'Corporations in the Age of Inequality.' *Harvard Business Review* (online). Retrieved May 1, 2018. (https://hbr.org/cover-story/2017/03/corporations-in-the-age-of-inequality).

Boyle, James. 2008. *The Public Domain: Enclosing the Commons of the Mind.* New Haven and London: Yale University Press. Available at: http://thepublicdomain.org/thepublicdomain1.pdf.

Brynjolfsson, Erik, and Andrew McAfee. 2016. *The Second Machine Age: Work, Progress, and Prosperity in the Time of Brilliant Technologies.* New York; London: W. W. Norton.

CIA. 2018. 'The World Factbook.' Retrieved May 1, 2018. (https://www.cia.gov/library/publications/the-world-factbook/docs/contributor_copyright.html)

Cohen, G. A. 1995. *Self-Ownership, Freedom, and Equality.* Cambridge: Cambridge University Press.

Creative Commons. 2018. 'When We Share, Everyone Wins.' Retrieved May 1, 2018 (https://creativecommons.org/).

Cwik, Bryan. 2016. 'Property Rights in Non-Rival Goods.' *Journal of Political Philosophy* 24(4), 470-86.

Dougherty, Tom. 2015. '*Yes Means Yes* : Consent as Communication.' *Philosophy & Public Affairs* 43 (3): 224–53.

Dworkin, Ronald. 2002. *Sovereign Virtue*. Cambridge, MA: Harvard University Press.

Economist. 2018. 'Free Exchange: The digital proletariat.' *Economist*. January 13, p. 69.

Englisch, Joachim. 2018. 'Digitalisation and the Future of National Tax Systems: Taxing Robots?' SSRN Paper 3244670. Retrieved 9 August 2019. https://papers.ssrn.com/abstract=3244670.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. 'Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks.' *Nature* 542 (7639): 115–18.

Fairfield, Joshua A. T. 2017. *Owned.* Cambridge: Cambridge University Press.

Frey, Carl Benedikt. 2019. *The Technology Trap: Capital, Labor, and Power in the Age of Automation*. Princeton, N. J.: Princeton University Press.

Hargittai, Eszter, and Gina Walejko. 2008. 'The Participation Divide: Content Creation and Sharing in the Digital age.' *Information, Communication & Society* 11 (2): 239–56.

Hess, Charlotte, and Elinor Ostrom. 2003. 'Ideas, Artifacts, and Facilities: Information as a Common-Pool Resource.' *Law and Contemporary Problems* 66: 111-45.

Himma, Kenneth Einar. 2005. 'Information and Intellectual Property Protection: Evaluating the Claim That Information Should Be Free.' *APA Newsletter on Philosophy and Law* 4: 3–9.

Hook, Leslie. 2018. 'Driverless Cars: Mapping the Trouble Ahead.' *Financial Times*, February 21, 2018.

Hughes, Chris. 2018. 'The Wealth of Our Collective Data Should Belong to All of Us'. *The Guardian*, 27 April 2018. Retrieved 9 August 2019, https://www.theguardian.com/commentisfree/2018/apr/27/chris-hughes-facebook-google-data-tax-regulation.

Kitchin, Rob. 2017. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Los Angeles: Sage.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge; New York: Cambridge University Press.

Lane, Julia I., Victoria Stodden, Stefan Bender, and Helen Fay Nissenbaum, eds. 2014. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. New York, NY: Cambridge University Press.

Lanier, Jaron. 2014. *Who Owns The Future?* London: Penguin.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. 'Deep Learning.' *Nature* 521 (7553): 436–44.

Lessig, Lawrence. 2004. *Free Culture*. New York: Penguin

Mayer-Schönberger, Victor, and Thomas Ramge. 2018. *Reinventing Capitalism in the Age of Big Data.* London: John Murray.

McKinsey Global Institute. 2015. 'Digital America: A Tale of the Haves and Have-Mores'.

Metz, Cade. 'In a Huge Breakthrough, Google's AI Beats A Top Player At The Game of Go', *Wired* 27th January 2016. Retrieved 12 October 2018. (http://www.wired.com/2016/01/in-a-huge-breakthrough-googles-ai-beats-a-top-player-at-the-game-of-go/)

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. 2015. 'Human-Level Control through Deep Reinforcement Learning.' *Nature* 518 (7540): 529–33.

Moore, Adam D. 2012. 'A Lockean Theory of Intellectual Property Revisited.' *San Diego Law Review* 49: 1069–1104.

Morozov, Evgeny. 2018. 'After the Facebook scandal it's time to base the digital economy on public v private ownership of data'. *The Guardian*. April 1, 2018. Retrieved October 2, 2018. (https://www.theguardian.com/technology/2018/mar/31/big-data-lie-exposed-simply-blaming-facebook-wont-fix-reclaim-private-information)

Nissenbaum, H. F. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Redwood City: Stanford University Press.

Nozick, Robert. 1974. *Anarchy, State and Utopia*. Oxford: Blackwell.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, London: Penguin.

Pollock, Rufus. 2009. 'The Economics of Public Sector Information.' *Cambridge Working Papers in Economics*. Retrieved October 2, 2018. (https://econpapers.repec.org/paper/camcamdae/0920.htm.)

Polson, Nicholas G, and James Scott. 2018. *AIQ: How People and Machines Are Smarter Together*. London: Bantam.

Posner, Eric A, and E. Glen Weyl. 2018. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton, NJ: Princeton University Press.

Rawls, John. 1999. *A Theory of Justice*. Rev. ed. Oxford: Oxford University Press.

Sadowski, Jathan. 2019. 'When data is capital: datafication, accumulation, and extraction.' *Big Data & Society* January-June 2019, 1-12.

Samuelson, Paul A. 1954. 'The Pure Theory of Public Expenditure.' *The Review of Economics and Statistics* 36 (4): 387.

Schradie, Jen. 2012. 'The Trend of Class, Race, and Ethnicity in Social Media Inequality.' *Information, Communication & Society* 15 (4): 555–71.

Silver, David., Hassabis, Demis. 'AlphaGo: Mastering the ancient game of Go with Machine Learning,' *Google Research Blog*, January 27th, 2016. Retrieved May 1, 2018. (https://research.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html.)

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, et al. 2017. 'Mastering the Game of Go without Human Knowledge.' *Nature* 550 (7676): 354–59.

Solove, Daniel J. 2012. 'Introduction: Privacy Self-Management and the Consent Dilemma Symposium: Privacy and Technology.' *Harvard Law Review* 126: 1880–1903.

Stiglitz, Joseph E. 1999. 'Knowledge as a Global Public Good.' In *Global Public Goods: International Cooperation in the 21st Century*, edited by Inge Kaul, Isabelle Grunberg, and Marc A. Stern, United Nat, 308–25. New York and Oxford: Oxford University Press.

Tegmark, Max. 2018. *Life 3.0: Being Human in the Age of Artificial Intelligence*. London: Penguin.

Urban, Tim. 2016. 'The AI Revolution: The Road to Superintelligence.' *Wait But Why*, January 22, 2015. Retrieved May 1, 2018. (http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html)

Weber, Steven. 2017. 'Data, development, and growth.' *Business and Politics* 19(3): 397-423

Wikipedia. 2018. 'Wikipedia: Copyrights.' Retrieved May 1, 2018. (https://en.wikipedia.org/wiki/Wikipedia:Copyrights).

Zhang, Xing-Zhou., Jing-Jie Liu and Zhi-Wei Xu. 2015. 'Tencent and Facebook Data Validate Metcalfe's Law.' *Journal of Computer Science and Technology* 30(2): 246-251.

Biographical Statements

Kai Spiekermann is Professor of Political Philosophy at London School of Economics, London, WC2A 2AE, UK.

Adam Slavny is Associate Professor at the University of Warwick, Coventry, CV4 7AL, UK.

David V. Axelsen is a Lecturer at the University of Essex, Colchester, CO4 3SQ, UK.

Holly Lawford-Smith is a Senior Lecturer in Political Philosophy at the University of Melbourne, Victoria 3010, Australia.

Revision Memo for Big Data Justice: A Case for Regulating the Global Information Commons

Dear Professor Levy,

Many thanks for tentatively accepting our manuscript for publication. We are again grateful for the detailed and generous comments provided by you and the referees. We have worked on improving the paper further by responding to these comments and by streamlining and ironing out small mistakes. We hope the new version finds your approval.

For your convenience, we have reproduced the reports below and added our responses with indent.

Editor

I admit that I am still puzzled about the demand for consent to unforeseen consequences in response to 5*, and I doubt that there is a generalizable principle at work there that can successfully distinguish this case from the Wilt Chamberlain example's complex downstream consequences of inequality. But I think this is a substantive disagreement on my part, not a problem that should delay acceptance of the manuscript.

> We have tried to emphasize our answer to this concern by making a change in the passage where we discuss the difference between Cohen's argument and ours. We now say that "Moreover, when Nozick's basketball fans pay to watch Wilt play, they are offering one small contribution to a much larger effect *of the same type*. When people contribute content to the information commons, however, they produce an entirely unrelated aggregate effect." (p. 21). In our view, what makes our setting structurally different is the discontinuity between contribution and result: if you pay money for a ticket, it is not very surprising that a potential outcome is more inequality (which is, of course, one viable line of argument to push back against Cohen). By contrast, if you contribute to the global information commons, more inequality is not something you would have anticipated. We believe there is a generalizable principle operating in the background, namely that whether implicit consent may be assumed depends on whether effects are foreseeable or unforeseeable, and whether they are just an aggregate outcome of many small actions, or an outcome that is qualitatively new.
> We acknowledge that this is a tricky debate. A fully satisfactory answer would need more space than is available in the present paper. We hope to return to this point in future publications.

Reviewer #2: Overall comments:

There are a few other scholars who make points that may be of interest to the authors that are not currently cited (though I leave it entirely up to them to consider whether or not these articles ultimately merit citation in their work):

On the point of the central value of data (public or private) in driving commercial value for companies and informing key decisions regarding business models:

o Jathan Sadowski, "When data is capital: Datafication, accumulation, and extraction," Big Data & Society, January-June 2019.

> Thank you for this helpful reference. We have now cited Sadowski on p. 8 and discuss the argument briefly in footnotes 15 and 18.

o Glen Weyl & Eric Posner, Radical Markets, is cited and discussed in the Data as Labor proposal, but can also be cited for the point regarding the need for high-quality data in developing AI

> Many thanks. We strengthened footnote 10, in which Posner & Weyl were already cited.

o MIT Technology Review puts out a number of reports published in conjunction with corporate partners on the value of data and the role it can or should play in driving business decisions.

> We struggled to identify directly relevant pieces in the MIT Technology Review, so we have refrained from referencing them in our paper.

Regarding the global dimension of data flows, the center/periphery phenomena of data use, and alternative conceptions from development theories of how to think of and respond to the inequality data may generate:
o Steven Weber, "Data, development, and growth," Business and Politics, 2017; 19(3): 397-423.

> Many thanks for this helpful pointer. We now cite Weber in footnote 12.

Regarding the argument that users retain certain rights to data after publication, additional engagement with/citation of Helen Nissenbaum's "Privacy in Context" may further strengthen the claim. This and related work by Nissenbaum may also strengthen the other arguments regarding use restrictions when use goes beyond norms.

> Reading Nissenbaum's book was very useful. We now cite her work on page 25 and discuss the implications briefly in footnote 21.

The Wilt Chamberlain example nicely makes the point regarding the indirect inequality effects of Free Use that ought to be taken into account. I am curious how the authors think of their focus on the distributive effects of data use, and their suggestion of a progressive data tax (both of which do seem focused on outputs), and their earlier framing of the focus on data as a focus on inputs. I do not think they two are contradictory, but perhaps worth making an explicit point that gaining clarity in the particulars of the input/production of wealth from data, lends insight into what kinds of output/distributional effects warrant attention and redress.

> In light of this comment we have lightly revised the passage on input-focused strategies on p. 4. In particular, we removed the stark contrast to output, as this invites the misunderstanding that we are not interested in outputs in the sense the referee reads the term. What interests us is to think more about the AI production process and its inputs.

Page specific comments:

Page 2:
- I think this draft establishes very well the central claim regarding the value of publicly-available data as a core driver of innovation. If the authors would like further citations from companies emphasizing the value of data, I would suggest Radical Markets by Posner & Weyl (Chapter 5 has a lot of language about the importance of high value data for developing AI). MIT Technology Review publishes a series of industry-aimed reports in conjunction with companies, one of which, "The Artificial Intelligence Imperative: Unlocking Data Insights to Fuel Business Growth and Innovation" nicely makes the point regarding how companies themselves recognize the centrality of data to ongoing competitiveness. However, neither explicitly makes the point regarding publicly available data, so authors should feel free to stick with the examples they have.

> We have made better use of Posner & Weyl, as explained above. We did not find that the MIT Technology Review is helpful to bolster our specific argument, as anticipated by the referee, so we decided not to use that item considering the strict page limit.

Page 3:
-"More specifically, we ask whether the large-scale use of data for the creation of proprietary artificial intelligence systems provides grounds for charging the owners of artificial intelligence."
I found this sentence to be important, but the use "charging" a bit unclear. Perhaps something like "More specifically, we ask whether the large-scale use of data for the creation of proprietary artificial intelligence systems provides grounds to [require payment] from the owners of artificial intelligence."

> We have made the change as suggested, many thanks.

Page 26:
-Perhaps this doesn't matter as much for the JOP audience, but in the final objection raised, for legal scholars it is worth clarifying that the authors mean moral rights, rather than legal rights, since the two can depart rather starkly on this case.

> We have inserted the qualifier “moral” once to avoid that misunderstanding.

Reviewer #3:

I am less satisfied, however, with the authors' response to my request for a deeper discussion of public goods and what makes them such. While there may be no one "definitive description" of what goods a generic public good provides, it does not follow that there is no answer to the question of what good the global information commons, as a public good, provides. I could think of several

examples, including the one listed in the previous paragraph of my review, namely, that it spurs technological progress to the benefit of society. The authors, too, were able to provide several options in their response, although they claim that the multiplicity of possible goods is a reason not to address them head-on. To my mind, however, understanding and articulating the actual goods being provided and to whom is key to forming a critical judgment about how to prioritize the global information commons. So I would have liked for them to have gone further in addressing this challenge than they did.

> While initially resisting this proposal, we have now decided to at least briefly take up this idea. To that effect, we added footnote 5, which states some ways the public good can be valuable: “as an educational resource, as a repository of human knowledge, a resource to spur technological innovation”.

***

> Many thanks again for the excellent comments provided, we have benefited a lot from this constructive critique and many excellent suggestions.
>
> Kai Spiekermann, Adam Slavny, David Axelsen, Holly Lawford-Smith